Final Project — Choose One Case

Each group selects **one** of the cases below. All follow the same pipeline:

- **Mandatory** Data Preprocessing + EDA
- **Mandatory** Supervised ML (Classification/Regression)
- **Mandatory** Unsupervised ML (Clustering + Dimensionality Reduction)
- **Mandatory** Deep Learning (DNN + Specialized Architecture)
- **Mandatory** Streamlit Deployment
- **Clear Dataset Specification**
- **No Optional Tasks** - All components are required and integrated

User Story 2: Customer Value Prediction & Segmentation

Data Preprocessing

As a data engineer at E-Commerce Pro
I want to process 500,000+ customer transactions into meaningful behavioral features
So that we build accurate customer lifetime value models

Dataset: Online Retail Dataset - UCI

- **Size:** 541,909 transactions, 8 features
- **Target:** `Customer Lifetime Value` (total customer spending)

Requirements:

- Clean data: remove cancelled orders (negative quantities, InvoiceNo starting with 'C')
- Handle missing CustomerIDs (approximately 135,000 records)
- Calculate Customer Lifetime Value: Total spending per customer (Quantity × UnitPrice)
- Create RFM features:
  - **Recency**: Days since last purchase per customer
  - **Frequency**: Total number of transactions per customer
  - **Monetary**: Total spending amount per customer
- Engineer temporal features: hour of day, day of week, month, season from InvoiceDate
- Handle categorical variables: Country, product categories from Description
- Normalize numerical features using StandardScaler

Exploratory Data Analysis

As a business analyst
I want to analyze 541,909 transactions to understand customer purchasing behavior
So that I can identify revenue optimization opportunities and customer patterns

Requirements:

- Customer distribution analysis by country (UK-focused business)
- Time-series analysis of sales trends across 2010-2011
- Customer lifetime value distribution and Pareto analysis (80/20 rule)
- RFM feature distributions and correlations
- Product popularity analysis using StockCode and Description
- Cohort analysis to track customer retention over time
- Seasonal patterns and peak purchasing periods
- Geographic analysis of customer spending patterns

Supervised ML

As a data scientist
I want to build regression models that predict customer lifetime value
So that we can identify high-value customers and optimize marketing strategies

Requirements:

- Implement Linear Regression with RFM features as baseline
- Build Random Forest regressor with 200 estimators for non-linear relationships
- Train XGBoost with hyperparameter tuning for optimal performance
- Evaluate models using RMSE, MAE, and R-squared metrics
- Perform feature importance analysis to identify key value drivers
- Use cross-validation to ensure model robustness and prevent overfitting
- Compare model performance against business benchmarks

Unsupervised ML

As a marketing strategist
I want to segment customers into meaningful behavioral groups
So that we can develop targeted marketing campaigns and personalized strategies

Requirements:

- Apply KMeans clustering (k=5) on RFM features to identify customer segments
- Use Gaussian Mixture Models for probabilistic cluster assignments
- Implement PCA for dimensionality reduction and cluster visualization
- Validate clusters using business metrics (average spend, purchase frequency, recency)
- Create detailed customer personas for each segment with characteristic patterns
- Analyze segment distribution across different countries and product categories
- Use silhouette scores to optimize number of clusters

Deep Learning

As a ML engineer
I want to develop deep learning models that capture complex customer behavior patterns
So that we improve prediction accuracy and discover hidden insights

Requirements:

- Build DNN regressor with 4 hidden layers (128, 64, 32, 16 neurons) for CLV prediction
- Implement Autoencoders for unsupervised customer feature learning
- Use cluster memberships as additional input features to DNN
- Apply batch normalization and dropout regularization (0.3)

- Compare deep learning performance against traditional ensemble methods
- Use embedding layers for high-cardinality categorical features
- Implement early stopping to prevent overfitting

Streamlit Deployment

As a product manager
I want an interactive customer analytics dashboard serving 541,909 transaction insights
So that marketing teams can make data-driven customer management decisions

Requirements:

- Customer lookup functionality with real-time lifetime value prediction
- Interactive segmentation explorer with cluster profiles and characteristics
- RFM analysis dashboard with customer scoring and filtering capabilities
- Campaign performance simulator with ROI projections by customer segment
- Export functionality for targeted marketing lists and customer segments
- Real-time visualization of customer behavior patterns and trends
- Model performance monitoring and prediction confidence scores

---